The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Learning Multimodal Word Representation via Dynamic Fusion Methods

Shaonan Wang,[1,2] **Jiajun Zhang,**[1,2] **Chengqing Zong**[1,2,3]
[1] National Laboratory of Pattern Recognition, CASIA, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China
{shaonan.wang,jjzhang,cqzong}@nlpr.ia.ac.cn

Abstract

Multimodal models have been proven to outperform text-based models on learning semantic word representations. Almost all previous multimodal models typically treat the representations from different modalities equally. However, it is obvious that information from different modalities contributes differently to the meaning of words. This motivates us to build a multimodal model that can dynamically fuse the semantic representations from different modalities according to different types of words. To that end, we propose three novel dynamic fusion methods to assign importance weights to each modality, in which weights are learned under the weak supervision of word association pairs. The extensive experiments have demonstrated that the proposed methods outperform strong unimodal baselines and state-of-the-art multimodal models.

Introduction

Representing the meaning of a word is a prerequisite to solve many natural language problems, such as calculating semantic relations between different words, finding the most relevant images of a word and so on. In recent years, computational semantic models that represent word meanings from patterns of word co-occurrence in corpora have received a lot of research interests (Turney and Pantel 2010; Mikolov et al. 2013; Clark 2015; Wang, Zhang, and Zong 2017a). However, compared to human semantic representation, these purely text-based models are severely impoverished for lacking perceptual information attached to the physical world. This observation has led to the development of multimodal word representation models that utilize both linguistic (e.g., text) and perceptual information (e.g., images, audios). Such models can learn better semantic word representations than text-based models, as evidenced by a range of evaluations (Andrews, Vigliocco, and Vinson 2009; Bruni, Tran, and Baroni 2014).

Learning good multimodal word representations relies not only on the quality of the word representations from linguistic and perceptual inputs, but also the ability to productively combine these representations. However, the existing multimodal models generally treat the word representations from different modalities equally. This is inconsistent with the fact that meaning of concrete words like `horse` and `computer` are mostly learned from perceptual experiences of seeing, touching and listening. In contrast, more abstract words, such as `hope` and `lovely`, are encoded mostly in linguistic modality rather than perceptual modality, which has been found in cognitive psychology (Wang et al. 2010; Binder et al. 2016) and computational experiments (Hill, Reichart, and Korhonen 2014; Hill and Korhonen 2014).

All these factors motivate us to build a multimodal model that can dynamically fuse information from linguistic and perceptual modalities according to different types of words. We can optimize the importance weights of different modalities for a word if the word has the gold representation. As no gold word representation exists in reality, we resort to word pairs which share the same meaning, so that they can guide each other. In this paper we utilize word association pairs[1], which are generated by subjects firstly reading a cue word and then writing down the first word(s) that come to mind. Some examples are `wealthy` and `rich`, `jigsaw` and `puzzle`, `larger` and `bigger`. We assume that these association word pairs can lead us to learn the importance weights for different modalities. For instance, representations of abstract words `larger` and `bigger` are composed by linguistic and perceptual vectors, and the linguistic vectors are more important in representing abstract word meaning (i.e., the two words share more similarity in linguistic modality). To achieve the goal of making these two association words obtain similar representations, the model will assign more weights to their linguistic vectors.

In light of these considerations, we propose three novel dynamic fusion methods to improve multimodal word representations. The three methods utilize a modality-specific gate, category-specific gate and sample-specific gate respectively, to learn different weights of linguistic and perceptual representations for each input modality, each supersense category and each word sample respectively. Furthermore, we perform extensive analysis to shed light on the principle of the proposed dynamic fusion methods. To summarize, our main contributions are two-fold:

- We present a novel dynamic fusion method for multi-

[1] We have also tried other resources, such as synonyms from WordNet. However, these datasets are noisy and perform slightly worse, thus we only report results of word associations.